Deciphering the Biological Mechanisms Driving the Phenotype of Interest

John Quackenbush Alejandro Quiroz Zárate

Deciphering the biological mechanisms driving the phenotype of interest

Abstract

The two key concepts of Neo-Darwinian evolution theory are genotype and phenotype. Genotype is defined as the genetic constitution of an organism and phenotype refers to the observable characteristics of that organism. Schematically the relationship between genotype and phenotype can be settled as

$$Genotype + Environment + Random\,Variation \xrightarrow{yields} Phenotype$$

This schematic representation has led to the fundamental problem of given the interactions of the genes and environment, up to what extent is possible to establish a relationship between gene structure and function to the phenotype (Weatherall, D. J., et. al., (2001)). Since R. A. Fisher establishing the basis of quantitative trait loci up to the work of Subramanian, et. al., (1995) gene set enrichment analysis, several statistical methods have been devoted to answer this question, some with more success and scientific repercussion than others.

In this work we attempt to answer to this question by delineating the biological mechanisms driven by the genes that are characterize the differences and actions of the phenotypes of interest. Our contribution resides on two pillars: we present an alternative way to conceive gene expression measurements and the use of functional gene set annotation systems as guided prior knowledge of the biological mechanisms that drive the phenotype of interest. Based on these two pillars we propose a method to infer the Functional Network Inference and an alternative method to perform expression Quantitative Trait Loci analysis. (eQTL) From the Functiona Network Inference method we are able to identify what mechanisms describe the behavior of most of the, there fore establishing its importance. The alternative method to perform eQTL analysis that we present, is more direct way to associated variations at a sequence level and

the biological mechanisms it affects. With this proposal we attempt to address two important issues of traditional eQTL analysis: statistical power and biological implications.

Table of Contents

Acknowledgements

I would like to acknowledge my friends and colleagues at Department of Biostatistics at Harvard University who have supported and helped me over the past few years. I would also like to thank all the extraordinary effort of Fundación México en Harvard and the Consejo Nacional de Ciencia y Tecnología. Without their support I would not have had the opportunity of this exciting endeavor.

I would like to especially thank my advisor, John Quackenbush, for his guidance, encouragement and constant support during the development of this dissertation. Throughout this time, John was not only a mentor but also a friend. His excitement and commitment to research has set a strong example that I hope to emulate.

I would like to thank Winston Hide and Curtis Huttenhower for many hours of invaluable discussions on the dissertation that helped me move forward. I also thank Benjamin Haibe-Kains for all his guidance during this time at John Quackenbush's lab.

A special thanks goes to Jess Mar for being an inspiring colleague and a friend.

Thanks to my fabulous wife Cristina and my beautiful daugther Julieta for all their patience and love throughout this extraordinary journey that has been my Ph.D. Everything is better with you girls beside me.

Finally, I would like to thank my parents and brothers. It is almost impossible to describe in words the endless support and unconditional love that they have given me on every step of the way. To you, *Papá, Mamá, Jesús y Rodrigo* I dedicate this work.

List of figures

Chapter 1: Introduction

There are two main concepts in Neo-Darwinian evolution theory: Genotype and Phenotype. Genotype refers to the all the genetic information that constitutes an organism. Phenotype refers to all the observable traits or characteristics of that organism. Considering gene expression as an observable trait, the main focus of all the reverse engineering methods is to model the complex relationships, functions and structure among the genes and elements it encodes. Over the past two decades the reduction of the costs to generate genomic data have increased its availability, enabling the possibility to develop reverse engineering methods to delineate the genetic architecture of gene regulation from gene expression measurements. With the aim to capture and model the complex behavior of the genes several statistical methods have been developed; the Linear Models for Microarray Data (LIMMA) (Smyth, G. K., et. al. (2002)); the Bayesian Estimation of Temporal Regulation (BETR) (Aryee, M., et. al. (2009)); Gene Set Enrichment Analysis (GSEA) (Mootha, V. K., et. al. (2003)); the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Dennis, G., et. al. (2003)). All these statistical methodologies test genes independently, ignoring their potential relationships aiming to infer the mechanisms that drive the disease of interest. Even though this gene-based approach had success, these methods often are not reproducible among independent datasets.

In this work we propose to make use of prior biological knowledge and the concept of repeated measurements of groups. As sources of prior biological knowledge throughout this work we mean functional gene set annotations such as REACTOME pathways (Croft, D., et. al., (2011)) or the Gene Ontology terms (Ashburner, M., et. al., (2000)). These functional annotations will be thought as gene groups throughout this work. The concept of integration of measurements will turn around the idea of genes thought as repeated measurements of such gene groups. The aim of this unification is to improve the robustness of the inferences by enhancing the

reproducibility of the results. In doing so we expect to obtain a better understanding of the mechanisms driving the disease of interest.

Based on the repeated measurement gene set approach we will focus in the development of three methods that have the goal to identify the mechanisms driving the disease of interest:

- A gene set selection method to identify the gene groups that differ most among the different stages of the disease under study making use of all the available information (genome wide scale) and treating genes as repeated observations of known biological classes.
- An algorithm to infer functional networks. In comparison to the existing gene set enrichment methods, our method attempts to identify biological classes that drive the behavior of distinct phenotypes of interest.
- A modification to the existing expression Quantitative Trait Loci (eQTL) methods. Our method attempts to identify significant associations between chromosomal regions and gene functional groups, exploiting the idea that its processes that are regulated rather than induced ones.

This work is divided in the following chapters: Chapter 2 presents the gene set selection method developed, along with an application and an analysis of its performance. Chapter 3 presents the method for functional network inference and exploits its application. Chapter 4 presents the alternative approach to the classical eQTL analysis: functional QTL (fQTL) and compares it with other methods.

Chapter 2: Gene set selection method

A wide range of statistical procedures for the analysis of DNA microarrays have been devoted to identifying sets of genes associated with biologically relevant phenotypes and which distinguish these from others (Golub, T. R., et. al., (1999), Van't Veer, L. J., et. al., (2002) and Wang, Y., et. al., (2005)). The standard approach is to first identify a *"significant gene list"* and then place these into a biological context by searching for enrichment of gene functional classes. The focus on biological functions is more informative than a simple *"significant gene list"*. This *"significant gene list"* paradigm, while widely used, has not been universally successful, given that most gene set-based expression signatures have not been validated in subsequent studies. There are many potential reasons for this, including the molecular heterogeneity of seemingly homogeneous biological samples, the sensitivity of gene expression measurements to small variations in the environment, the fact that genes themselves do not operate in isolation but rather function as elements of complex pathways which themselves are more important and informative than individual genes, and the robustness of the methods we use to select genes and classify samples.

Here we propose an alternative approach to the *"significant gene list"* paradigm based on the concept of repeated measurements of gene sets. Our objective is to provide robust inferences, in the sense of being replicable among independent datasets. By doing, so we hope to obtain a better understanding of the biological processes driving the differences between phenotypes. The proposed approach is based on Bayesian statistical modeling methods.

In this chapter we will present the proposed methodology together with an application to a public accessible breast cancer dataset. We will examine the performance and robustness of our method by analyzing two independent public accessible Breast cancer datasets. Finally we present the conclusions of the proposed methodology.

2.1 The methodology

Our goal is to identify biological functions that describe in a concise way the mechanisms that differentiate phenotype states. We will use the Reactome pathways of the Gene Ontology (GO) to define functional gene groups. We will use a linear model based on the mean expression of these gene groups to compare the behavior between phenotypes. The linear model assumption allows analysis of datasets collected using a wide variety of exploratory design; from cross-sectional disease comparisons to time series analysis. To control for multiple comparisons we will use the sparsity priors proposed by Carvalho, C. M. et. al., (2008). These sparsity priors have been successfully applied in a wide variety of scenarios: from denoising musical audio (Févotte, C., et. al., (2008)) to genomics.

As an example, consider a time series experiment in which there are two stages of the disease of interest, **A** and **B.** There is a total of **J** samples and an irregular time course sampling with **H** time points. We assume there are **I** gene groups and a total of **K** genes being assayed. In this scenario it is possible that genes can be in several gene groups. Let $\boldsymbol{i}$ denote the gene group, $\boldsymbol{j}$ the sample, $\boldsymbol{k_i}$ the total number of genes (repeated measurements) associated with gene group $\boldsymbol{i}$ and $\boldsymbol{t_h}$ a point of the time course. For each sample $\boldsymbol{j}$, at each time point, we measure the mean gene expression level *y* for gene group $\boldsymbol{i}$, and assume that these measurements are normally distributed:

$$(y_{ij}^{t_1}, y_{ij}^{t_2}, \ldots, y_{ij}^{t_H}) \sim N_H\left(\boldsymbol{\mu_i}, \frac{\Sigma_i}{k_i}\right)$$

The model of the mean for each gene group *i* is:

$$\boldsymbol{\mu_i} = \boldsymbol{\beta_i} + \boldsymbol{\alpha_i^*} 1(stage = A),$$

where $\boldsymbol{\beta_i} = \left(\beta_i^1, \beta_i^2, \ldots, \beta_i^H\right)$. Here $\boldsymbol{\beta_i}$ is the average expression level of gene group $\boldsymbol{i}$ across time points. And $\boldsymbol{\alpha_i^*}$ captures the disease-time interaction: the average time expression level

differences between the stages of the disease of interest. Biologically, $\boldsymbol{\alpha_i^*}$ captures two behaviors of time series data: trajectories that between stages of the disease present a non-parallel behavior (NPB) which are crucial in the understanding of the biological mechanisms that drive a differential response and the trajectories that between stages of the disease present a parallel behavior (PB), which represent a similar response only distinctive by intensity across time between stages. To model (NPB) trajectories, $\boldsymbol{\alpha_i^*}$ is parameterized as:

$$\begin{aligned}\boldsymbol{\alpha_i^*} &= \boldsymbol{\alpha_i^0} + \boldsymbol{\alpha_i} \\ &= \left(\alpha_i^0, \alpha_i^0, \cdots, \alpha_i^0\right) + \left(0, \alpha_i^1, \cdots, \alpha_i^{H-1}\right) \\ &= \left(\alpha_i^0, \alpha_i^0 + \alpha_i^1, \cdots, \alpha_{i+}^0 \alpha_i^{H-1}\right)\end{aligned}$$

where, $\boldsymbol{\alpha_i^0}$ captures a baseline difference in expression between stages of disease across time and the $\boldsymbol{\alpha_i}$ characterize the NPB of the trajectories. On the other hand we can model all the differential trajectories (PB and NPB) with the following parameterization of $\boldsymbol{\alpha_i^*}$

$$\boldsymbol{\alpha_i^*} = \boldsymbol{\alpha_i} = \left(\alpha_i^1, \cdots, \alpha_i^H\right).$$

On both parameterizations it is $\boldsymbol{\alpha_i}$ the parameter of interest. As mentioned in the beginning of this section the adjustment for multiple comparisons is performed through the sparsity priors (Carvalho, C. M., et. al., (2008)) on $\boldsymbol{\alpha_i}$:

$$\boldsymbol{\alpha_i} \sim \pi_i^{\alpha} \boldsymbol{N}_H(0, \Sigma_{\alpha}) + (1 - \pi_i^{\alpha}) \delta_0(\boldsymbol{\alpha_i}),$$

$$\pi_i^{\alpha} \sim \rho^{\alpha} Beta\big(am, a(1-m)\big) + (1 - \rho^{\alpha}) \delta_0(\pi_i^{\alpha}),$$

$$\rho^{\alpha} \sim Beta(sr, s(1-r)).$$

Here, π_i^{α} is the probability of $\boldsymbol{\alpha_i}$ taking a nonzero value. This parameter controls the sparsity behavior. In this context sparsity means that many of these probabilities will be small or 0 and

only a few of these will take on large values: the gene groups with strongest differential behavior between the stages of the disease. The Beta prior distribution for nonzero values of π_i^{α} will be fairly diffuse favoring large probabilities. The Beta prior distribution for ρ^{α} is chosen to favor very small values. The reasoning behind this selection of prior distributions is that a large fraction of gene groups are expected to have uncoordinated behavior among its repeated gene expression measurements. Thus a large fraction of π_i^{α} are expected to be concentrated around 0 and only a small number with high values. For this purpose the parameters of the Beta prior distribution for π_i^{α} and ρ^{α} are chose to have a high mean and a diffuse variability and a small mean with diffuse variability respectively. The prior distribution for the variance in the model, Σ_{α} and Σ_i will follow and inverse Wishart distribution in an effort to maintain close form expression on the full conditional distributions. This enables to estimate the parameters of interest ($\boldsymbol{\alpha_i}$) through a Gibb's sampler. The selection of gene groups with differential behavior between the stages of the disease is based on the parameter:

$$\hat{\pi}_i^{\alpha} = P(\boldsymbol{\alpha_i} \neq 0)$$

Our focus will be to select gene groups with high values of $\hat{\pi}_i^{\alpha}$. It has to be highlighted that the gene group selection done by the proposed method have by design, more power than the ones obtained by the combined statistical procedures that constitute a classical gene set enrichment analysis.

2.2 Application

The proposed method was applied on a dataset with cross-sectional design. Gene expression measurements taken from fresh-frozen tumor samples from patients with lymph-node-negative breast cancer who were treated during 1980-95 (Minn, A. J., et. al., (2007)), for whom Estrogen receptor status (ER+ or ER-) was recorded. Estrogens are hormones, which means that they function as signaling molecules. Estrogens control multiple physiological processes

including growth, differentiation and function of reproductive system. Paradoxically, estrogen can be both a beneficial and a harmful molecule. Unlike normal breast cells, cancer cells arising in the breast do not always have receptors for estrogen. Breast cancers that do have estrogen receptors are said to be *"estrogen receptor-positive"*. In these cancer cells, growth is under control of estrogens and generally has a better prognosis. In contrast, those breast cancers that do not possess estrogen receptors are *"estrogen receptor-negative"*. For this, the growth is not governed by estrogen. These cancers are more aggressive and unresponsive to anti-estrogens (Rochefort, H., et. al., (2003)).

Min and colleagues (Minn, A. J., et. al., (2007)) used Affymetrix U133A GeneChips to profile gene expression in 286 fresh-frozen tumor samples from patients with lymph-node-negative breast cancer who were treated during 1980-95, but who did not receive systemic neoadjuvant or adjuvant therapy (Minn). These samples correspond from the data set used in (Wang, Y., et. al., (2005)) with GEO reference accession number GSE2034, from the tumor bank at the Erasmus Medical Center in Rotterdam, Netherlands. An additional 58 estrogen receptor-negative samples were added from (Minn, A. J., et. al., (2007)) GEO (GSE5327). In total 209 tumor samples are classified as ER+ and 135 as ER-. Even though this data set comes from a 5-year follow-up design, we treat the data as cross-section given our interest to analyze ER status.

2.2.1 The model under a cross-sectional data design

The model for gene group selection we used

$$y_{ij,T} \sim N\left(\mu_i, \frac{\sigma_i^2}{k_i}\right),$$

where $y_{ij,T}$ stands for the mean gene group measurement for group *i,* patient *j*, *T* for the ER status of the patient (positive or negative), and k_i the total number of genes associated with group *i.* Here

we assume the variance is different between groups. We define the linear model that describes the mean gene group behavior by

$$\mu_i = \beta_i + \alpha_i^* \times 1(T = ER-).$$

In this model β_i is the overall average expression level across ER status for the gene group i. And α_i^* captures the differential expression between ER status for gene group *i*. Selection of the gene groups relies on the parameter α_i. This parameter captures the possible effects ER status and is the parameter of interest. For this analysis, Molecular Function (MF) of GO by the Molecular Signature Database (MSigDB) at the Broad Institute will be used as source of gene groups ise used. Only gene groups with more than four elements were considered on the analysis. In total 396 MFs were considered. In Figure 2.1 it is presented the MFs identified with a probability of at least 90% of being differentially expressed between ER status. It can be seen that the MFs related to *acetyltransferase* are a consistent finding related to ER status by the activity of gene NAT1. NAT1 transcript is reported to be good prognostic markers of ER+ status (Wakefield, L., et. al., (2008)). Besides this interesting finding, NAT1 is also reported to have an important role in breast cancer progression (Wakefield, L., et. al., (2008)).

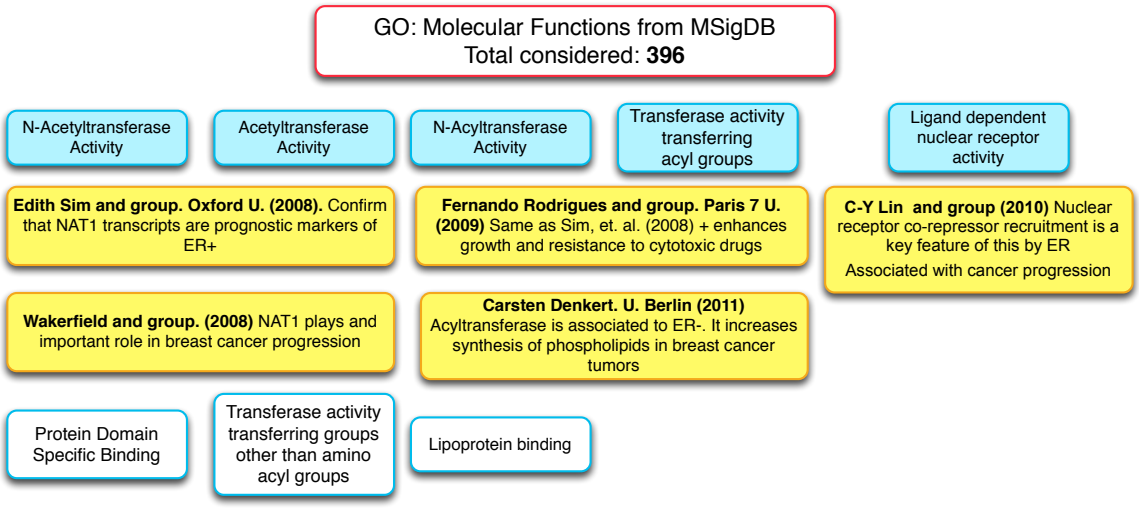


Figure 2.1 Molecular functions (MF) selected by the proposed method with a probability of 90% of being differentially expressed between ER status are the ones in blue. The MFs highlighted in blue are MFs that had a reported relationship with Breast cancer and ER status. Its association is explained in yellow.

The MFs related to *acyltransferase* are a consistent finding too with ER status in Breast cancer. It has been reported that *actyltrasnferase activity* enhances growth and resistant of tumors to cytotoxic drugs (Ragunathan, N.,et al., (2009)). Its activity also increases the synthesis of phospholipids in Breast cancer tumors (Brockmöller, S. F., et. al., (2012)), making the cellular membrane of Breast cancer tumor cells more resistant to target drugs. Finally the MF *ligand dependent nuclear receptor activity* is also consistent to ER activity in Breast cancer. It has been reported that nuclear receptor co-repressor recruitment is a key feature of ER, which is related to Breast cancer progression (Merrell, K. W, et. al., (2011)).

2.3 Analysis of the performance of the proposed methodology

The objective of the proposed methodology is to provide robust inferences. To provide a measure of robustness we evaluated the proposed methodology based on concordance and signal replicability of the selected groups. By concordance we refer the replicability of such inferences

among independent datasets. By signal replicability of the selected groups we refer to selecting the gene groups with highest signal among independent datasets.

For the performance analysis of the inferences obtained by the proposed methodology two independent datasets were analyzed: TRANSBIG data (TRANSBIG) (Desmedt, C., et. al. (2007)) consisting of 198 frozen samples with 134 tumor samples classified as ER+ and 64 as ER- and the Schmidt M, et al. (2008) (Schmidt) dataset consisting of 200 lymph node-negative Breast cancer samples with 162 tumor samples classified as ER+ and 38 as ER-.

To obtain an objective insight of the performance of the proposed methodology, we provide a comparison against a widely used gene set selection approach: the combination of LIMMA to establish a significant gene list with a cut-off of 5% of False Discovery Rate (FDR) and then GSEA for the enrichment analysis with a cut off of 5% of Bonferroni correction. We divide the gene groups enriched from GSEA into GSEA-L and GSEA-U, to refer to the enriched significant gene groups from the lower part of the distribution of the test statistic or from the upper part of the distribution of the test statistic respectively.

2.3.1 Concordance performance

The results of the comparison are shown in Figure 2.2. The objective of this comparison is to see how concordant are the MFs identified by our proposed methodology and a widely used methodology for gene set enrichment analysis among different datasets. From the p-values shown on Figure 2.2 we see that the inferences obtained from the three methods between these three independent datasets are concordant. Even though having an acceptable performance, the p-value measure is sensitive to the total sample size.

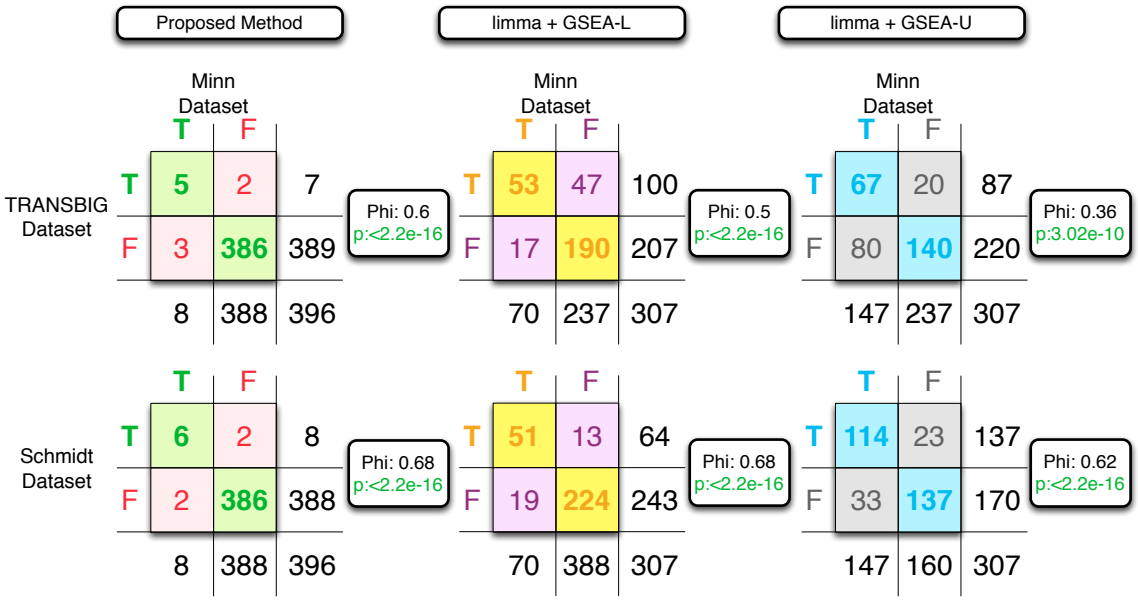


Figure 2.2 Concordance results from the proposed method and from GSEA-L and GSEA-U. This figure shows the concordance of the MFs identified using the Minn dataset as reference. In all the contingency tables **T** stands for *"Differentially identified gene group"* and **F** stands for *"Not differentially identified gene group"*

In this case the proposed method considered 396 gene groups and both GSEA-L and GSEA-U considered 307 gene groups only. To address this issue, we used the statistic Phi for a more objective measure. The Phi statistic is just the square root of the test statistic of the contingency table divided by the sample size. The Phi statistic ranges from -1 to 1. Having Phi set to -1 means that there is an inverse association between the variables test on the contingency table. Having Phi set 0 means there exist no association at all. And finally having Phi set to 1, means there is a strong positive association among the variables tests on the contingency table. From the Phi statistic we see that on the comparison of TRANSBIG vs Minn datasets, our proposed method does better. On the other hand on the comparison of Schmidt vs Minn datasets all three methods have a comparable performance.

2.3.2 Signal among independent datasets

To analyze the signal of the selected gene groups we establish the following as a signal measure

$$Signal = \left| \frac{\mu_{ER-} - \mu_{ER+}}{\sigma_{(ER-,ER+)}} \right|.$$

The reasoning is the following: measure the difference in means between ER positive and negative weighted by the pooled standard deviation for each gene group. The absolute value is taken just for plotting presentation purposes. This measure provides an empirical way to objectively analyze the behavior in means between ER status for each gene group. The Minn dataset is established as the reference dataset given it has more sample size, enabling a more reliable overview of the difference between ER status.

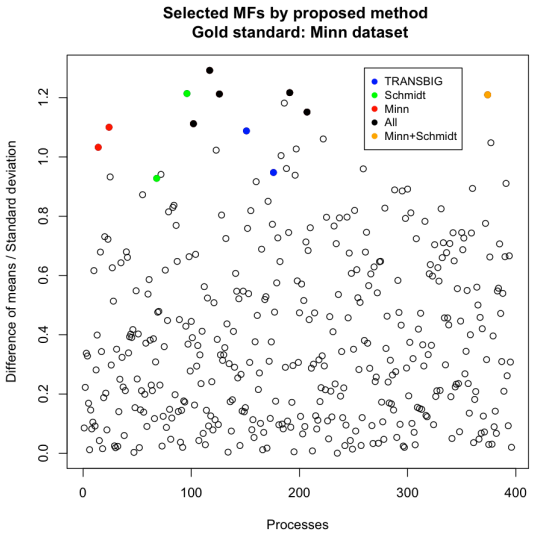


Figure 2.3 Scatter plot showing the signal of each gene group based on the Minn dataset. The selected gene groups were identified by the proposed method.

In Figure 2.3 we can see that, relative to the information of the Minn dataset, the identified gene groups by the proposed method are within the MFs with the highest signal.

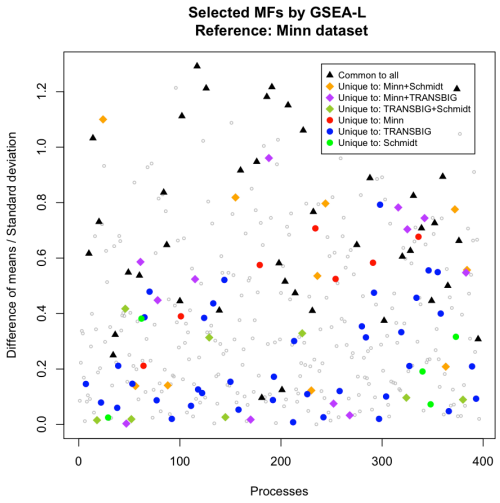


Figure 2.4 Scatter plot showing the signal of each gene group based on the Minn dataset. The selected gene groups were identified by GSEA-L.

In Figure 2.4 we can see the behavior of the gene groups identified by GSEA-L relative to the information of the Minn dataset. Even though the identified gene groups by GSEA-L are within the MFs with the highest signal, GSEA-L also selects gene groups with low signal. This is not a desired characteristic: identifying gene groups with uncoordinated behavior among its repeated gene expression measurements is something any gene set selection method should avoid.

In Figure 2.5 we can see the behavior of the gene groups identified by GSEA-U relative to the information of the Minn dataset. The performance of the selected gene groups is not ideal with respect to signal. There is a high number of gene groups with very small signal, which is not a desired characteristic.

From Figures 2.3, 2.4 and 2.5 we see that the proposed method selects gene groups that are among the ones with highest signal. On the other hand the gene groups selected by GSEA-L and GSEA-U have also high signal, but along with this desired property on the selected gene groups there are a high number of gene groups that do not have signal at all.

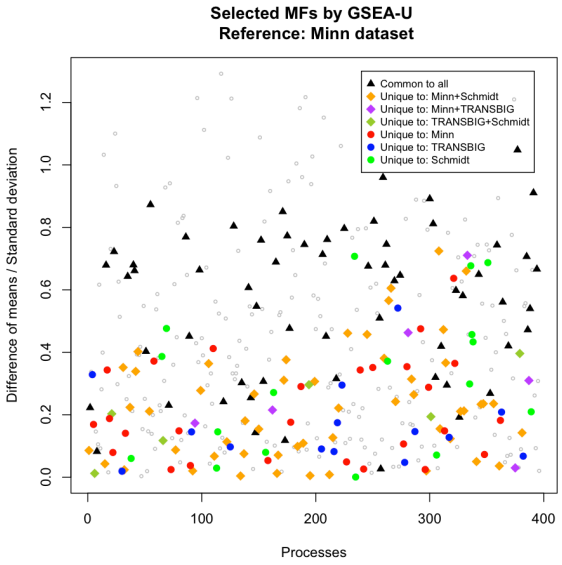


Figure 2.5 Scatter plot showing the signal of each gene group based on the Minn dataset. The selected gene groups were identified by GSEA-U

2.4 Discussion

In this chapter we have shown the method we propose for gene set selection. This method is based on the proposed unification of prior biological knowledge and the concept of integration of measurements. The selection is based on the assumption that a linear model is suitable to capture changes between the different stages of the disease of interest. Although simplistic this assumption, it enable us to apply our method in wide variety of data designs, from the very common cross-sectional to the time course data design.

The method was applied to Breast cancer dataset with cross-sectional data design. The interest of this application was to identify the Molecular Functions from GO that presented a different behavior between ER status. From the eight MFs identified by the proposed method, five are consistent with what is known of ER status in Breast cancer. From these MFs selected we have learned that the main theme that characterizes the differences between ER status is a

differential in prognosis and the cellular functions related to drug resistance. Both of these themes are consistent with what is known of ER status and Breast cancer.

The objective of the proposed method was to provide robust inferences; inferences that can be replicable among independent datasets. To measure the robustness two characteristics were analyzed: concordance and signal maintenance among independent datasets. The performance of the proposed method was compared against a widely used methodology: LIMMA for the significant gene list discovery and GSEA for the enrichment analysis. Based on concordance, our method has a comparable performance. On the other hand the proposed method has an outstanding performance in identifying gene sets with high signal among independent datasets. From the results provided we are able to conclude that are method provide useful results. The inferences made by the proposed method replicable across independent data sets and are consistent with the biology known with the disease of interest.

Chapter 3: Biological functional networks

The philosophy of gene set enrichment methods is to provide biological context to a significant gene list. The objective is to infer the mechanisms that characterize the differences between the different stages of the disease of interest. From the proposal of DAVID (Dennis, G., et. al. (2003)) through the successful GSEA method (Subramanian, A., et. al. (2005)), gene set enrichment analysis provides only a list of functional classes. This list of functional classes although insightful is somewhat incomplete, in the sense that we aren't able to identify which classes are crucial to activate/drive the disease under study or which ones even though significant, have secondary roles. To address this problem we prose a modification to a class of algorithms that are design for network inference on large sparse datasets (datasets with high number of variables and a small number of samples): the Sparse Candidate Algorithms, proposed by Friedman, N., et. al. (1999). Our modification enables the possibility to use the information of gene expression measurements to guide the potential links between the gene groups or functional classes. Our iterative method provides a set of networks from which we will resume their topological properties in order to infer what we define as biological functional networks. We provide a new perspective to analyze gene sets that are significantly related to the stages of the disease of interest and at the same time attempt to identify the sets that have crucial roles in understanding the mechanisms driving the phenotype of interest.

In this chapter we present our method for functional network inference. We present the iterative procedure. We also present an application to a public accessible Breast cancer data set. The performance of the proposed method is measured based on the concordance of the inferences obtained.

3.1 The method

The objective of the proposed work is to analyze how the molecular interactions influence the associations between biological functions and whether these associations delineate meaningful mechanism with respect to the different stages of the disease of interest. We refer to biological functions as predefined gene set annotations from databases like Reactome pathways (Croft, D., et. al., (2011)) or the Gene Ontology (Ashburner, M., et. al., (2000)). The associations between these functional classes will be estimated through a linear model. These associations will be the links of the networks that constitute the output of our method: a set of *functional networks*.

The algorithm for functional network estimation is described for a cross-sectional data design framework. A straightforward modification is needed in the case of a time series data design.

3.1.1 The algorithm

We will explain in detail the algorithm for functional network inference

3.1.1.1 The setting

Let $FC = \{fc_1, fc_2, \ldots, fc_m\}$ be the set of functional classes from the biological annotation set of interest where each functional class, $fc_i = \{g_{i1}, g_{i2}, \ldots, g_{in}\}$, consists of genes g_{il}. Because we are conceiving that genes are repeated measurements of the functional classes (gene groups), throughout this work we will conceive the observations of the functional classes as the mean expression of the genes within the respective functional class (gene group) k fc_k. So each fc_k will be a vector of dimension $n \times 1$ where n is the number of samples. The same dimension will hold for the observations of the genes g_i. It will be assumed that no functional classes will have the same exact set of genes and that these classes will have a minimum number of genes, k. To infer the associations between functional classes a linear model is assumed:

$$fc_i = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} X_{ij}, \qquad (3.1)$$

where X_{ij} belongs to the set *FC*. The interest of this work is not to estimate the parameters of the regression, given that these have a meaningless interpretation. Our interest is to identify the variables that take place in the regression, to perform variable selection. This problem is much more attainable than to estimate regression parameters. Given the sparsity of the data (more functional classes –variables- than samples) it is not practical to apply statistical techniques that are *ex professo* for this ill data setting that identify *a* solution to (3.1). Instead, the focus of the proposed algorithm is to provide a "variety" of solutions that are the most likely predictors for each functional class.

The algorithm to infer the variables $\{X_{ij}\}_{j \neq i}$ is based on a class of algorithms proposed by Friedman, N., et. al., (1999) to infer gene networks: the Sparse Candidate Algorithm. These algorithms restrict the possible parents for each gene. Then by making use of a statistical score the algorithm attempts to maximize such score by making use of an exhaustive search for the best parents within the restricted parent set for each gene. Such procedure presents a crucial drawback; it can potentially lead to local maximum. This because once the candidate parents for each variable, the algorithm is committed to them. In this work it is proposed to incorporate random sampling in the maximization of the statistical score in the selection of the best parents instead of the exhaustive search. By this, the proposed procedure has the ability to avoid any local maximum.

3.1.1.2 The algorithm

The iterative procedure of the propose algorithm is divided in two stages: the gene stage and the functional class stage.

At the gene stage, the update of the network is divided in two steps: the individual step and the group step.

- At the individual step, for each gene, g_i, we will focus on the genes that belong to the functional classes that serve as predictors of the functional classes for which gene g_i belongs to (understanding that a gene may be play a function in several biological mechanisms). The process of selecting candidate genes to predict the behavior of another is equivalent as searching for the best network that fits the data. This is known to be an NP-hard problem. To simplify this search we follow the guidelines of Zhu, J., et al. (2004): restrict the number of candidate predictors up to P_0. So every gene is considered, except gene g_i, (only when the experimental data design is a cross-sectional study) and randomly a decision is made to select P_0 candidate genes weighting them by the number of appearances in the functional classes that serve as predictors. Setting the g_i as the objective, the score of all the possible regression models that can be obtained from these P_0 predictors is compared against the score of the regression model with the actual predictors. The predictors of the regression model with the maximum score are selected as proposals.
- The update at the group level is performed taking into account the overall score of the predictors that are proposed against the overall score of the actual predictors, through random sampling.

At the functional class stage the update of the network is dived in two steps, as in the gene stage: the individual step and the group step.

- At the individual step, for each functional class fc_i, we will focus on the functional classes that are associated to the genes that serve as predictors of the genes that belong to the functional class fc_i. The process of selecting the candidate functional classes is the same

as in the individual step at the gene stage, with just a subtle modification. Because some biological functions will share a large amount of their genes, the addition of either one to the model as predictor would tend to be indistinguishable. This is the reason to be of the *"Similarity Step"*. A set of candidate predictors is discarded if there exist at least one among the P_0 selected, that is has a similarity measure above a threshold chosen by the user with at least one of the actual predictors. As a similarity measure between sets Dice's coefficient was chosen (Equation 3.2):

$$DC = \frac{2|X \cap Y|}{|X| + |Y|}, \qquad (3.2)$$

where X and Y are sets.

- The update at the group level is the same as at the gene stage.

The proposed algorithm is an iterative procedure that is run for a total of *no.sim* number of simulations. To avoid any bias respect to the starting networks a burn in period is established of *burn.in* initial iterations. The output of the algorithm is a set of networks, $\mathbb{N}^{FC} = \{N_i^{FC}\}: |\mathbb{N}^{FC}| = no.sim - burn.in$ where $N_i^{FC} = \{(fc_k, X_{kj})\}$ for which fc_k is the functional class objective and X_{kj} is a functional class that serve as predictor of fc_k. The links/associations of interest, (fc_k, X_{kj}), are the within the whole set for networks, $\mathbb{N}^{FC}$, that satisfy:

$$P\left((fc_i, X_{ij}): \exists N_j^{FC} \in \mathbb{N}^{FC}, (fc_i, X_{ij}) \in N_j^{FC}\right) \geq c \qquad (3.3)$$

for any c between *0* and *1*. These associations, (fc_k, X_{kj}), that satisfy (3.3) form a *resume functional network*: the ultimate goal of the algorithm. The associations/links from the *resume functional network* are crucial in identifying the functional classes that act as predictors of others. Thus the *"key players"* of this network are the functional classes that can predict or describe the

behavior of most of the other functional classes. It will be these classes the ones that will be of interest in our inferences. An outline of the functional network algorithm is presented in Figure 3.1.

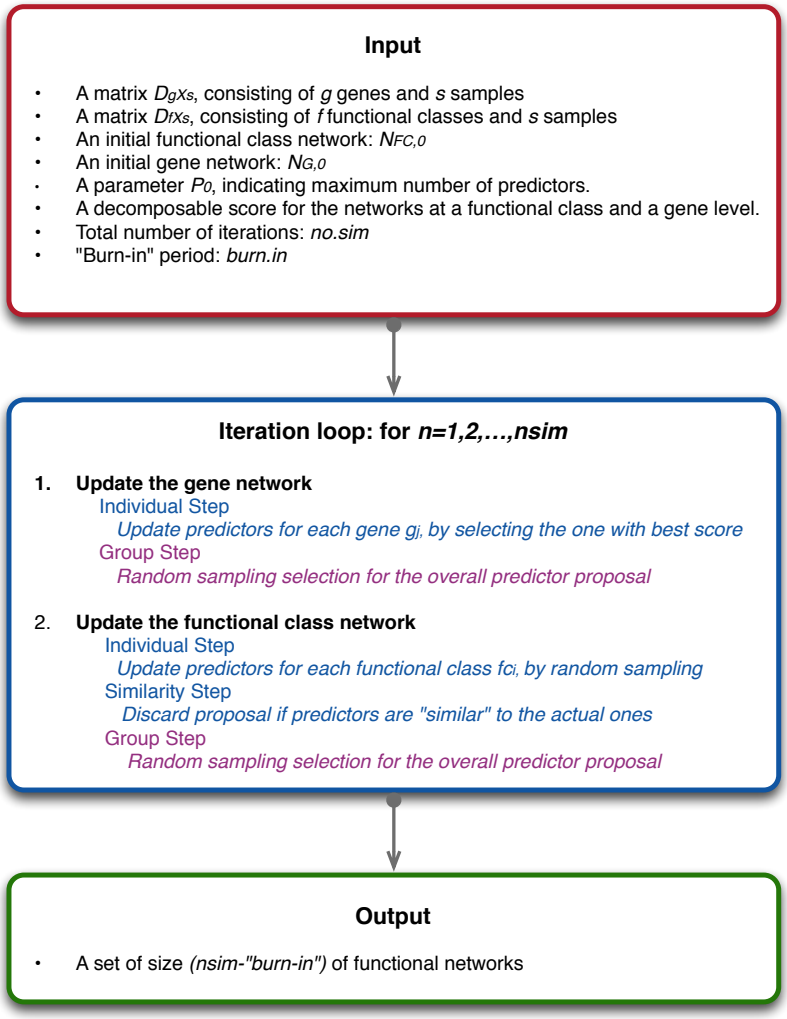


Figure 3.1 Outline of the functional network algorithm. We describe the input of the method, each step of the iterative procedure and the output generated.

3.2 Application

The functional network algorithm was applied on a dataset with cross-sectional design. It consists of gene expression measurements from Breast cancer patients taken from frozen archival

tumor samples from the TRANSBIG study used on Desmedt, C., et. al. (2007) for whom Estrogen receptor status (ER+ or ER-) was recorded.

Estrogens control multiple physiological processes including growth, differentiation and function of reproductive system. Paradoxically, it can also function as a harmful molecule. Unlike normal breast cells, cancer cells arising in the breast do not always have receptors for estrogen. Breast cancers that do have estrogen receptors are said to be *"estrogen receptor-positive"*. In these cancer cells, growth is under control of estrogens and generally has a better prognosis. In contrast, those breast cancers that do not possess estrogen receptors are *"estrogen receptor-negative"*. In this case, the growth is not governed by estrogen. These cancers are more aggressive and unresponsive to anti-estrogens (Rochefort, H., et. al., (2003)).

Desmedt, C., et. al.,(2007) profiled gene expression measurements from 198 frozen archived tumor samples from patients with lymph-node-negative breast cancer clinical diagnosed between 1980 and 1998 (GEO ID: GSE7390). From the 198 tumor samples 134 are classified as ER+ and 64 as ER-. Even though this data set comes from a 13.6 median-year follow-up design, the way the data is conceived for this analysis was cross-sectional.

3.2.1 Functional network inference for a cross-sectional design

The proposed algorithm is applied to infer a functional map within each ER status. For this purpose the model proposed by equation (3.1) was applied. To learn the links of the resume functional network within the patients with ER – and within ER+ only the significantly differentiated genes between ER status are considered. Reactome pathways with at least 10 genes were selected as source of functional annotation from the MSigDB. The objective of the resume networks is to identify the Reactome pathways that are the key players within each ER status. A total of 3 predictors in the linear model are considered at the gene and functional stage and as a cut off of Dice's similarity measure (Equation 3.2) 87.5% was chosen. As a decomposable score

the Residual Sums of Squares was chosen to analyze the convergence of the algorithm. The algorithm was run for a total of 6,000,000 iterations with a *burn in* period of 5,000,000. In Figure 3.2 we see the stable behavior reached within the selected range of iterations of the total sum of squares of the whole resume network within ER- and ER+.

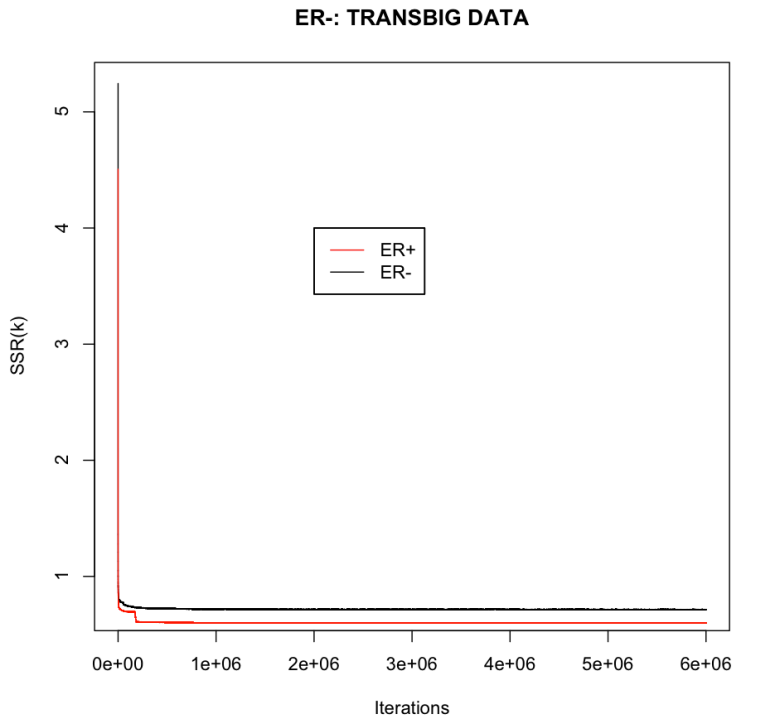


Figure 3.2 Plot of the behavior of the Residual sum of squares (SSR) for ER+ (red) and ER- (black) for the TRANSBIG dataset.

To obtain the Reactome pathways that characterize ER- or ER+ we will focus on each of the resume functional networks and obtain the key pathways based on the following measure:

$$PEC = \frac{Number\ of\ liks\ of\ the\ node}{Maximum\ number\ of\ links\ from\ all\ nodes\ in\ network} \qquad (3.4)$$

The pathways obtained using this measure defined as the Proportional Edge Count (PEC), are pathways involved describing the behavior of other pathways. These pathways will have the notion *"driving"* the activity within the respective ER status. By setting a threshold of 0.8 for the

PEC measure seven Reactome pathways were identified as key players in ER+. Among these five are related to ER+ and Breast cancer: *"p53 independent DNA damage response", "REV mediated nuclear export of HIV1 RNA", "RNA pol II CTD phosphorylation and interactions with CE", "TrkA signaling from the plasma membrane"* and *"Formation of early elongation complex"*. For example, *"p53 independent DNA damage response"* is associated with the fact that p53 is a target of estrogen receptor alpha, modulating DNA damage and growth suppression in ER+. The pathway *"Formation and early elongation complex"* and *"RNA pol II CTD phosphorylation and interaction with CE"* are related in the process of expression of the proto-oncogene MYB. MYB is expressed in the majority of ER+ breast tumors and is controlled by ERα and a transcriptional pausing mechanism involving an attenuation site located downstream from its transcription start site. ERα binds close to this site and drives transcription beyond this attenuation site for the complete synthesis of the transcripts. In the binding process of ERα the P-TEFb complex (CDK9/Cyclin T1) is recruited resulting in an increase RNA polymerase II phosphorylation (Mitra, P., et. al., (2012)). Finally is has been reported that estrogen treatment results in an increased expression of trkA ((McMillan, P. J., et. al., (1996)), (Sohrabji, F., et. al., (1995))), associating *"TrkA signaling from the plasma membrane"* with ER+ in Breast cancer. On the other hand by setting a threshold of 0.8 for the PEC measure eleven Reactome pathways were identified as key players. Among these 5 are related to ER- and Breast cancer: *"Formation of early elongation complex", "Cell surface interactions at the vascular wall", "Glucose and other sugar SLC transporters", "Glucose transport" and "Immunoregulatory interactions between a lymphoid and a non lymphoid cell"*. For example *"Formation of early elongation complex"* is related to ER status and Breast cancer through the activity of the oncogene MYB and as we have seen is expressed in the majority of ER+ breast tumors and not in ER- breast tumors. The pathways *"Glucose and other sugar SLC transporters"* and *"Glucose transport"* are also associated with ER- breast cancer. It has been reported that in locally invasive primary breast cancer, ER- negative tumors display higher (18)F-FDG than ER+ ones. Gene expression data

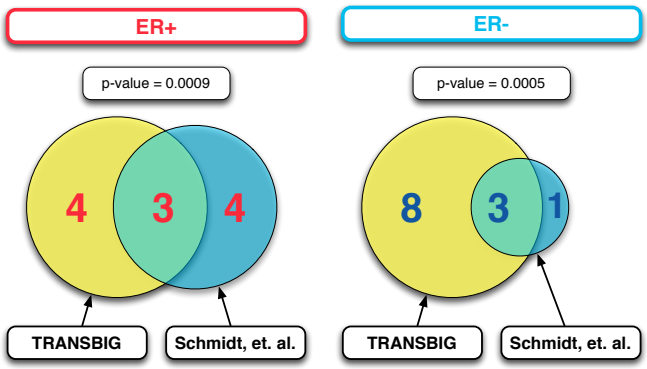
confirms and identifies genes associated with increase glucose. These findings and the fact that patients with higher glycemic load a higher risk of ER- breast cancer than the ones without suggest an association of insulin and ER- breast tumor growth. The pathway *"Immunoregulatory interaction between a lymphoid and non lymphoid cell"* has also association in ER- breast tumors. It has been reported that Lymphoid stroma and comedo-necrosis correlated with higher tumor grade, which appears to have a significant correlation with ER- tumors.

3.3 Method comparison

The objective of the proposed algorithm is to provide robust inferences. To provide a measure of robustness the proposed methodology was evaluated based on concordance. By concordance we refer the replicability of such inferences among independent datasets.

For the performance analysis of the inferences obtained by the proposed methodology an independent datasets was analyzed: the Schmidt M, et al. (2008) (Schmidt) dataset consisting of 200 lymph node-negative Breast cancer samples with 162 tumor samples classified as ER+ and 38 as ER-.

In Figure 3.3 we present the results of the concordance analysis. We compared the key players identified on each independent data set. On this analysis the key players are defined as Reactome pathways that had a PEC measure (Equation 3.4) greater than 0.8 on the networks generated on ER+ and ER- status. To provide an objective measure of whether the concordance was significant or not, we provide a p-value for comparisons made for each ER status. This p-value is generated through a Fisher exact test. By this we take into consideration the number of Reactome pathways that are expected and the total number of pathways considered. We see that for ER+ and ER- the concordance is significant. Given that this is conclusion is made through a Fisher p-value we can reject the hypothesis that the Reactome pathway selections under both datasets is not driven by the effect of ER status.



$$PEC = \frac{\#links(node)}{\max\{\#links(node)\}} > 0.8$$

Figure 3.3 Results of the concordance analysis from the algorithm for functional network inference. The numbers in bold on both Venn diagrams are the number of Reactome pathways that had a PEC measure (Equation 3.4) greater than 0.8. The p-values are for the each Venn diagram.

3.4 Discussion

In this chapter we present a novel method that attempts to identify gene functional classes that can delineate the potential biological mechanisms that characterize phenotype of interest. This iterative algorithm is based on a class of algorithms to infer gene networks: Sparse Candidate Algorithms (Friedman, N., et. al., (1999)). These algorithms provide a unique solution to the network estimation for sparse data sets (small number of samples and high number of features).

With the inclusion of the random sampling, our method has the opportunity of exploring the whole predictor space, with the ability of not getting *"trapped"* in a local maximum/minimum. Based on the unification of prior biological knowledge and the concept of genes as repeated measurements, we provide a method to infer networks at a group level by exploiting the idea that genes within the same biological mechanism or gene group potentially share regulators that perform similar functions.

The proposed method was applied to Breast cancer dataset with cross-sectional design. The analysis was based on Reactome pathways. From the inferences presented we are able to see that the Reactome pathways identified as key players through the PEC measure (Equation 3.4) are consistent with reported knowledge with ER status and Breast cancer.

The objective of the proposed method is to provide robust inferences: inferences that can be replicable among independent datasets. We analyzed the performance of robustness of the inferences through the concordance of Reactome pathways identified as key players between independent Breast cancer datasets. Based on the results obtained comparing the results between the TRANSBIG and Schmidt dataset we see that we have the proposed method presents inferences that are significantly concordant among independent datasets.

In conclusion, this work presents a method that attempts to identify the key processes that can help to understand how the molecular interactions can characterize the differences at a phenotype level. From the results provided along with the analysis of the performance we can conclude that the method provide useful inferences, and that these inferences are also consistent with the biology known of the disease of interest.

Chapter 4: Functional QTL analysis

The focus of this work relies in understanding the complex architecture of gene expression. Our effort focuses in understanding from gene expression measurements the complex processes at the genetic level. Microarray gene expression measurements have been used in wide variety of contexts, from the classical differential expression analysis up to the classification of tissues for disease diagnosis. With the objective in understanding the complex process of gene regulation, gene expression measurements have been conceived as continuous trait in the analysis of Quantitative Trait Locus analysis. These analyzes aim to identify the chromosomal region, the DNA polymorphism (binding site or a regulatory protein -transcription factor-) that affect the levels of gene expression in an inheritable way. This statistical tool is been denominated eQTL analysis.

Traditionally eQTL analysis associates genetic variation to gene expression measurements and from there attempt to assign biological function. Even though this technique has had several successes and has provided an insight into what can be proxy for regulatory associations, the association to biological function lacks of significance meaning given that it involves the combination of statistical procedures.

In this work we present an alternative approach for eQTL analysis. Our approach is based on the unification of prior biological knowledge as sources for gene group definitions and the concept of integration of measurements. Our model associates directly genetic variation to biological function by conceiving genes as repeated measurements of the gene groups considered from the biological annotation chosen.

In this Chapter we present our proposed method for eQTL defined as functional Quantitative Trait Loci analysis: fQTL analysis. We also present an application to an Alzheimer's disease dataset. And finally we present the ongoing work to complete this proposal.

4.1 The method

The objective of the proposed work is to identify the associations between variations at a genetic level and biological functions, based on the concept of genes acting as repeated measurements of such biological functions (gene groups). We refer to biological functions as predefined gene set annotations from databases like Reactome pathways (Croft, D., et. al., (2011)) or the Gene Ontology (Ashburner, M., et. al., (2000)). The associations will be estimated through a linear model. These associations provide an insight of the chromosome regions that affect the biological functions associated with the disease of interest.

We describe in detail the model for the fQTL analysis method proposed. There are a total of **K** subjects, **J** single nucleotide polymorphisms (SNPs), **I** gene groups and **H** genes. Let *k* denoted the subject, *j* the SNP, *i* the gene group and h_i for the gene within gene group *i*. The fQTL model is

$$y_{ijkh_i} = \beta_{0ij} + \beta_{1ij}\left(SNP{:}\,ADD_j\right) + \beta_{2ij}1(Phe = 1) + \beta_{3ij}\left(SNP{:}\,ADD_j\right) * 1(Phe = 1)$$

$$\varepsilon_{ijkh_i} \sim N\left(0, \sigma_{ij}^2\right).$$

From this setting it is immediate to see that our model by design will capture associations between SNPs and biological functions with more power than a traditional eQTL analysis would. Our main interest is to identify the associations between SNPs and gene groups for which the effect differs among stages of the disease under study. This effect relies in β_{3ij}, which are the parameters of interest.

4.2 Application

In this work we provide an fQTL analysis on a cohort of control and late-onset Alzheimer disease (LOAD) human brain samples provided by Webster, J. A., et. al. (2009). Raw genotype data and normalized gene expression data is provided. For our analysis we preprocessed the raw

genotype data, excluding genotypes with less than 90% in sample call rate and in SNP call rate. We also excluded genotypes that rejected the Hardy-Weinberg test to 0.05 significant level. Finally we excluded the genotypes that had less than five subjects for each allele class (cero, one or two minor alleles). In the end our analysis considered 208,631 SNPS, 363 subjects, 187 controls and 176 LOAD. Gene expression measurements we normalized with the rank invariant transformation.

According to the National Institute of Aging *Alzheimer's disease is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks.* This disease is among the most common cause of dementia among older people, affecting an approximate 5.1 million people in America.

For this analysis we used the Molecular Functions (MFs) provided by GO. Only the MFs with more than 7 elements were considered. Our interest is on the associations between SNPs and gene groups for which the effect differs among stages of the disease under study (β_{3ij}). The significant effects were identified using a Bonferroni significance level of 0.05. Figure 4.1 presents a composite Manhattan plot showing all the SNPs presenting an association with a MF for which there is an effect between Alzheimer status. From this Manhattan plot we are able to see that there are twenty-two MF identified. From these eleven MFs consistent with what is known about the mechanisms associated to Alzheimer (Figure 4.2). These elven MFs provide a wide overview of the potential mechanisms that are affected by sequence variations. From the activity of *small GTPase regulator activity* that enhances the generation of Beta amyloid and synaptic dysfunction up to *Vitamin binding*, that is associated with hormone receptors in patients with Alzheimer's disease.

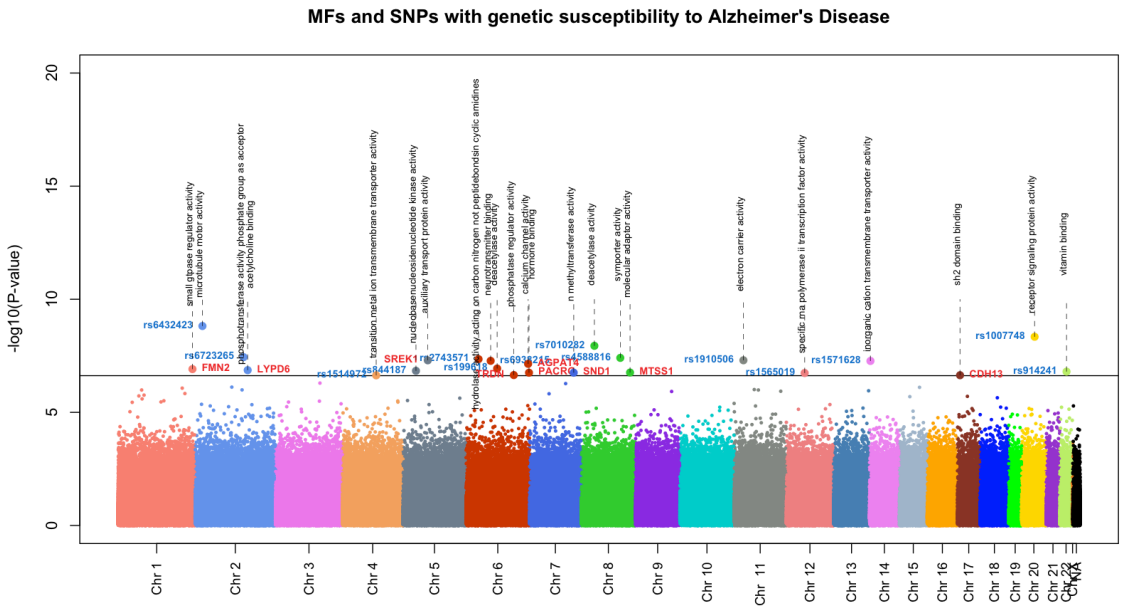


Figure 4.1 Composite Manhattan plot, showing the locations of the SNPs and the MFs that differ in effect between Alzheimer status.

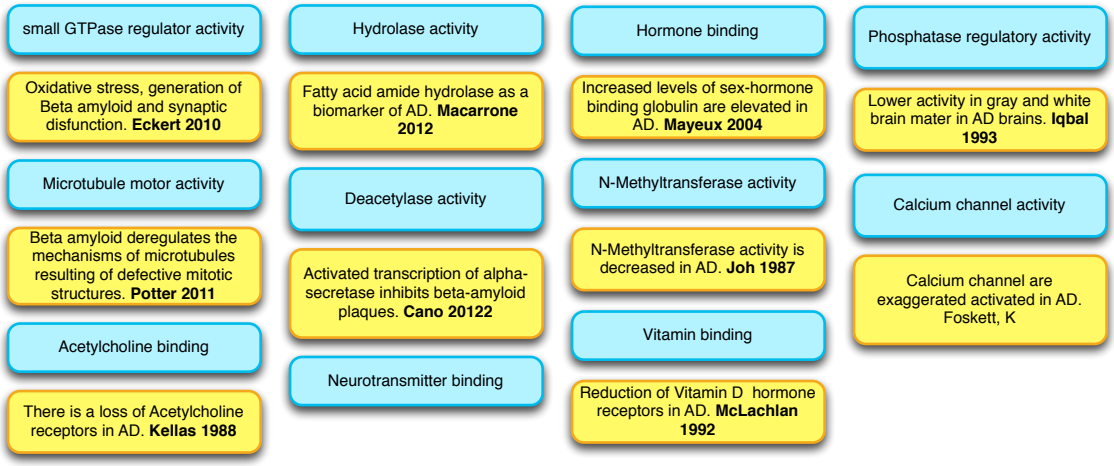


Figure 4.2 Eleven from the twenty-two molecular functions (MFs) having a significant association with a SNP for which there is a differential effect of Alzheimer status. These MFs are highlighted in blue. Highlighted in yellow are the literature findings reporting the association for each MF to Alzheimer.

4.3 Method comparison

The inferences obtained from the proposed fQTL analysis are compared to a classical eQTL analysis. We performed a classical eQTL analysis on the same set of genes considered for the fQTL analysis and the same linear model. By using the same Bonferroni significance level of 0.05 a list of genes were identified as being associated to a SNP with a differential effect to Alzheimer status. This list of significant genes was enriched using DAVID from the National Institute of Allergy and Infectious Diseases (NIAID-NIH). From this enrichment analysis a total of twenty-six Molecular functions were identified using a Bonferroni correction of

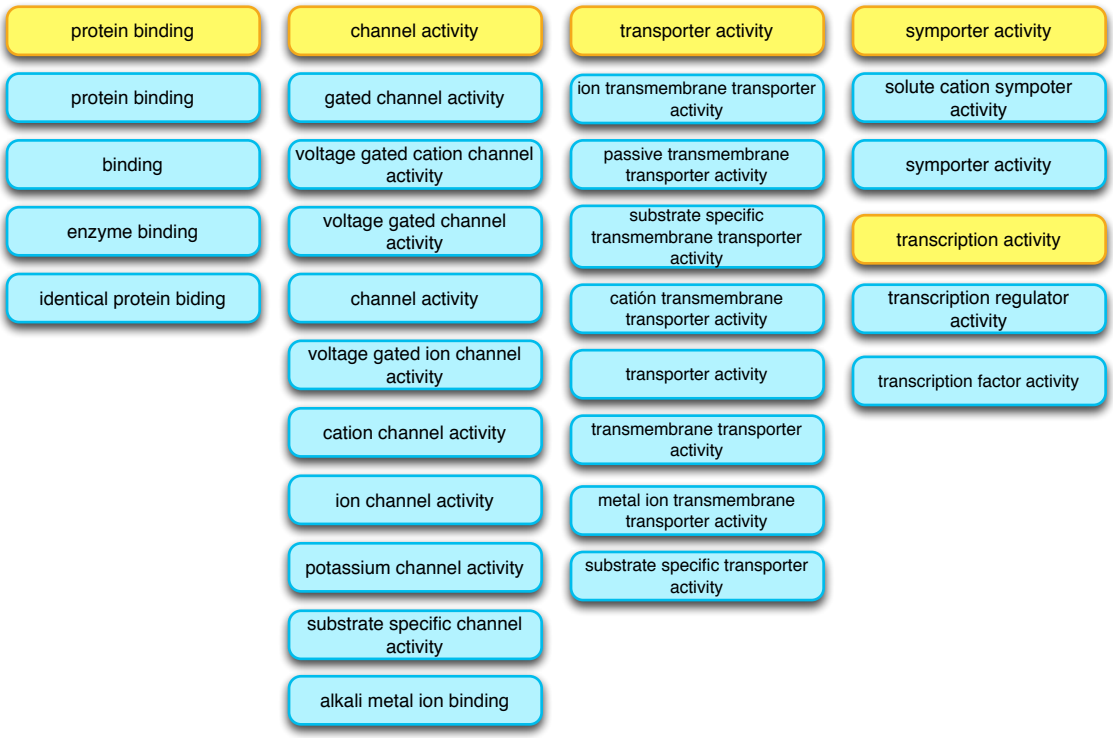


Figure 4.3 Molecular functions (MFs) identified using DAVID with a Bonferroni significance level of 0.05. This enrichment analysis was done using the gene list provided from the classical eQTL analysis performed. Highlighted in blue are the MFs identified and in yellow the classification themes that clustered them.

5%. In Figure 4.3 we show these MFs identified from DAVID's enrichment analysis. Even though there are more MFs identified their activity can be resumed in five major classes: *protein*

binding, channel activity, transporter activity, symporter activity and transcription activity. In comparison with the MFs identified by the fQTL analysis, our inferences provide a broader and consistent view of the mechanisms that are driving by these genetic variations. It has to be highlighted that our associations, by design, have more power than the ones obtained by combining a classical eQTL and a widely used gene set enrichment tool like DAVID.

4.4 Discussion and future work

In this chapter we have shown the method we propose for Quantitative Trait Loci analysis. This method is based on the proposed unification of prior biological knowledge and the concept of integration of measurements. This method is a variation to the classical eQTL analysis, attempting to associate directly chromosomal regions to biological functions that are affected by the disease status. By considering genes as repeated measurements of gene groups, by design, our approach has power in identifying such associations than the ones obtained by combining a classical eQTL and a widely used gene set enrichment tool like DAVID.

The method was applied to cohort of postmortem samples with late-onset Alzheimer disease. The interest of this application was to identify the Molecular Functions from GO that presented an association with SNPs for which their group behavior is affected by the Alzheimer status. Our method presents a broader overview of the MFs that are influenced by the interaction of SNPs and Alzheimer status, than a traditional eQTL analysis.

As continuation to this work, we propose to analyze an independent dataset in order to analyze the performance of the proposed method in terms of concordance: replicability of the inferences obtained between independent datasets. Besides this direction we pretend to analyze whether the inferred statistical associations provide biological insight, by analyzing the regulatory proteins (transcription factors -TFs- or microRNAs) between SNPs and its associated biological function. Being able to establish this confirmation at the biological level will enable the

possibility to provide insightful and useful biological hypothesis between functions associated to the disease of interest and at is potential regulatory mechanisms.

Bibliography

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2009). *Essential Cell Biology*. Garland Science, 3rd Edition.

Anurag, G., Leong, D. T., Fen, B. H., Brat, S. S., Thiam-Chye, L. and Werner, H. D. (2007). Osteo-maturation of adipose-derived stem cells required the combined action of vitamin D3, β-glycerophosphate, and ascorbic acid. *Biochemical and Biophysical Research Communications*. **362**, 17-24.

Aryee, M., Gutierrez-Pabello, J., Kramnik, I., Maiti, T and Quackenbush, J. (2009) An Improved Empirical Bayes Approach to Estimating Differential Gene Expression in Microarray Time-course Data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics*, **10**(1), 409

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubind, G. M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature* Genetics. **25**(1): 25-29.

Bonofiglio, D., Gabriele, S., Aquila, S., Catalano, S., Gentile, M., Middea, E., Giordano, F. and Ando, S. (2005) Estrogen receptor alpha bins to peroxisome proliferator activated receptor response element and negatively interferes with peroxisome proliferator-activated receptor gamma signaling in breast cancer cells. *Clinical Cancer Research*. **11**(17), 6139-6147.

Brockmöller, S. F., Bucher, E., Müller, B. M., Budczies, J., Hilvo, M., Griffin, J. L., Orešič, M., Kallioniemi, O., Iljin, K., Loibl, S., Darb-Esfahani, S., Sinn, B. V., Klauschen, F., Prinzler, J., Bangemann, N., Ismaeel, F., Fiehn, O., Dietel, M and Denkert, C. (2012). Integration of Metabolomics and Expression of Glycerol-3-phosphate Acyltransferase (GPAM) in Breast Cancer-Link to Patient Survival, Hormone Receptor Status, and Metabolic Profiling. *Journal of Proteome Research*.**11**, 850-860.

Buck, M. B. and Knabbe, C. (2006). TGF-Beta Signaling in Breast Cancer. *Annals of the New York Academy of Sciences, Estrogens and Human Diseases.* **1089**, 119-126.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., West, M. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. Journal of the American Statistical Association. **103**(484), 1438-1456.

Callihan, P., Mumaw, J., Machacek, D. W., Stice, S. L. and Hooks, S. B. (2011). Regulation of Stem Cell Pluripotency and Differentiation by G Protein Coupled Receptor. *Pharmacology and Therapeutics.* **129**, 290-306.

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*. **227**, 561-563.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P. and Stain, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research.* **39**, D691-D697.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*. **4**:R60.

Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, Viale G, Delorenzi M, Zhang Y, d'Assignies MS, Bergh J, Lidereau R, Ellis P, Harris AL, Klijn JGM, Foekens JA, Cardoso F, Piccart MJ, Buyse M and Sotiriou C. (2007). Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series. *Clinical Cancer Research.*. **13**(11), 3207-3214.

Edwards, J. C. W. (2000). Fibroblast Biology Development and Differentiation of Synovial Fibroblasts in Arthritis. *Arthritis Research.* **2**, 344-347.

Fan, T. M., Barger A. M., Sprandel, I. T. and Fredrickson R. L. (2008). Investigating TrkA Expression in Canine Appendicular Osteosarcoma. *Journal of Veterinary Internal Medicine*. **22** (5), 1181-1188.

Févotte, C. Torrésani, B., Daudet, L., and Godsill, S. J. (2008). Sparse Linear Regression with Structured Priors and Application to Denoising of Musical Audio. *IEEE Transactions on Audio, Speech and Language Processing.* **16**(1), 174-185.

Filardo, E. J. (2002). Epidermal Growth Factor Receptor (EGFR) Transactivation by Estrogen via the G-protein-coupled Receptor, GPCR-30: A Novel Signaling Pathway with Potential Significance for Breast Cancer. *Journal of Steroid Biochemistry and Molecular Biology.* **80**, 231-238.

Friedman, N., Nachman, I. and Peér, D. (1999). Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.* 206-215.

Fung, T. T., Hu, F. B., McCullough, M. L., Newby P. K., Willet W. C. and Holmes M. D. (2006). Diet Quality is Associated with the Risk of Estrogen Receptor-Negative Breast Cancer in Postmenopausal Women. *Nutritional Epidemiology.* **136**, 466-472.

Gardner, T. S. and Faith J. J. (2005). Reverse-engineering Transcription Control Networks. *Physics of Life Reviews*. **2**, 65-88.

Gery, S., Virk, R. K., Chumakov, K., Yu A. and Koeffler H. P. (2007). The Clock Gene PER2 Links the Circadian System to the Estrogen Receptor. *Oncogene.* **26**(57), 7916-7920.

Golub, T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S. Molecular Classification of Cancer: Class Discovery and Class Prediction by Expression Monitoring. (1999). *Science*. **286**: 531-7.

Harla, L. C., Coates, R. J., Block, G., Greenberg, R. S., Ershow, A., Forman, M., Austin, D. F., Chen V., and Heymsfield, S. B. (1993). Estrogen Receptor Status and Dietary Intakes in Breast Cancer Patients. *Epidemiology*. **4**(1), 25-31.

Hausman, G. J. and Dodson M. V. (2012). Stromal Vascular Cells and Adipogenesis: Cells within Adipose Depots Regulate Adipogenesis. *Journal of Genomics.* **1**, 56-66.

Hilvo, M., Denkert, C., Lehtinen, L., Müller, B., Brockmöller, S., Seppänen-Laakso, T., Budczies, J., Bucher, E., Yetukuri, L., Castillo, S., Berg, E., Nygren, H., Sysi-Aho, M., Griffin, J. L., Fiehn, O., Loibl, S., Richter-Ehrenstein, C., Radke, C., Hyötyläinen, T., Kallioniemi, O., Lijin, K. and Oresic, M. (2011). Novel Theranostic Opportunities Offered by Characterization of Altered Membrane Lipid Metabolism in Breast Cancer Progression. *Cancer Research*. **71**(9), 3236-3245.

Ishunina, T. A., van Heerikhuize, J. J., Ravid R. and Swaab D. F. (2003). Estrogen receptors and metabolic activity in the human tuberomamillary nucleus: changes in relation to sex aging and Alzheimer's disease. *Brian Research.* **988**(1-2), 84-96.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Älgenäs, C., Lundeberg, J., Mann, M. and Uhlen, M. (2010). Defining the Transcriptome and Proteome in Three Functionally Different Human Cell Lines. *Molecular Systems Biology*. **6**, 450.

McMillan, P. J., Singer, C. A. and Dorsa, D. M. (1996). The Effects of Ovariectomy and Estrogen Replacement on trkA and Choline Acetyltransferase mRNA Expression in the Basal Forebrain of the Adult Female Sprague-Dawley Rat. *The Journal of Neuroscience.* **16**(5), 1860-1865.

Merrell, K. W., Crofts, J. D., Smith, R. L., Sin, J. H., Kmetzsch K. E., Merell, A., Miguel, R. O., Candelaria, N. R. and Lin C. Y. (2011). Differential Recruitment of Nuclear Receptor Coregulators in Ligand-Dependent Transcriptional Repression by Estrogen Receptor-a. *Oncogene*. **30**, 1608-1614.

Minn, A. J., Gupta, G. P., Padua, D., Bos, P., Nguyen, D. X., Nuyten, D., Kreike, B., Zhang, Y., Wang, Y., Ishwaran, H., Foekens, J. A., Van de Vijver, M. and Massagué, J. (2007) Lung Metastasis Genes Couple Breast Tumor Size and Metastatic Spread. *PNAS.* **104**(16), 6740-6745

Mitra, P., Pereira, L. A., Drabsh, Y., Ramsay, R. G. and Gonda, T. J. (2012). Estrogen receptor-α recruits P-TEFb to overcome transcriptional pausing in intron 1 of the MYB gene. *Nucleic Acids Research.* **40**(13), 5988-6000.

Mootha, V.K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. and Groop, L. C. (2003). PGC-I alpha-responsive gene involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics.* **34**, 267-273.

Muller, P., Parmigiani, G., and Rice, K. (2007). FDR and Bayesian Multiple Comparison Rules. *Bayesian Statistics.* **8**, 349-370.

Ragunathan, N., Dairou, J., Pluvinage, B., Martins, M., Dupret, J. M. and Rodrigues-Lima F. (2009). Human Arylamine N-Acetyltransferase 1 (NAT1) as a Target of Chemotherapeutic Drugs in Breast Cancer: Cisplatin as Model. *Molecular and Cellular Pharmacology.* **1**(1), 7-10.

Rochefort, H., Glondu, M., Sahla, M. E., Platet, N. and Garcia, M. (2003). How to target estrogen receptor-negative breast cancer? *Endocrine-Related Cancer*. **10**, 261-266.

Sanchez, A. M., Flamini, M. I., Baldacci, C., Goglia, L., Genazzani, A. R. and Simoncini, T. (2010). Estrogen Receptor-a Promotes Breast Cancer Cell Motility and Invasion via Focal Adhesion Kinase and N-WASP. *Molecular Endocrinology.* **24**(11), 2114-2125.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*. **270**, 467-470.

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H. A., Hengstler, J. G., Kölbl, H. and Gehrman, M. (2008). The Humoral Immune System Has a Key Prognostic Impact in the Node-Negative Breast Cancer. *Cancer Research.* 68(13): 5405-5413.

Shang, Y., Hu, X., DiRenzo, J., Lazar, M. A. and Brown, M. (2000). Cofactor Dynamics and Sufficiency in Estrogen Receptor-Regulated Transcription. *Cell.* **103**, 843-852.

Simonini, P. S. R., Breiling, A., Gupta, N., Malekpour, M., Youns, M., Omranipour, R., Malekpour, F., Volina, S., Croce, C. M., Najmabadi, H., Diederichs, S., Sahin, Ö., Mayer, F., Hoheisel, J. D. and Riazalhosseini, Y. (2010). Epigenetically Deregulated microRNA-375 is Involved in a Positive Feedback Loop with Estrogen Receptor α in Breast Cancer Cells. *Cancer Research.* **70**, 9175-9184.

Smyth, G. K., Ritchie, M., Thorne, N., Wettenhall, J. and Shi, W. Linear Models for Microarray Data. (2002). Bioconductor package. Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia.

Sohrabji, F., Miranda, R. C. G. and Toran-Allerand, D. (1995). Identification of a Putative Estrogen Response Element in the Gene Encoding Brain-Derived Neurotrophic Factor. PNAS. **92**, 11110-11114

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy S. L., Golub, T. R., Lander E. S. and Mesirov, J. P. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *PNAS*. **102**(43), 15545-15550.

Van 't Veer, L. J., Dai H., Van de Vijver, M. J., He Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., Van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. Â. and Friend, S. H. (2002). Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature.* 415,530-536.

Wakefield, L., Robinson, J., Long, H., Ibbitt, J. C., Cooke, S., Hurst, H. C. and Sim E. (2008). Arylamine N-acetyltransferase 1 Expression in Breast Cancer Cell Lines: A Potential Marker in Estrogen Receptor-Positive Tumors. *Genes, Chromosomes and Cancer*. **47**(2), 118-126.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yanh, F., Talantov, D., Timmermans, M., Gelder, M. E. M. G., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D. and Foekens, J. A. (2005). Gene-expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer. *Lancet*. **365**, 671-79.

Webster, J. A., Gibbs, J. R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., McCorquodale D. S. III, Cuello, C., Leung, D., Bryden, L., Nath, P., Zismann, V. L., Joshipura, K., Huentelmn, M. J., Hu-Lince, D., Coon, K. D., Craig, D. W., Pearson, J. V., NACC-Neurophatology Group, Heward, C. B., Reiman, E. M., Stephan, D., Hardy, J. and Myers, A. J. (2009). Genetic Control of Human Brain Transcript Expression in Alzheimer Disease. *The American Journal of Human Genetics.* **84**, 445-458

Wollbold J, Huber R, Pohlers D, Koczan D, Guthke R, Kinne RW and Gausmann U. (2009). Adapted Boolean Network Models for Extracellular Matrix Formation. *BMC Systems Biology*. **3**, (77).

Yang, H. J., Xia, Y. Y., Wang, L., Liu, R., Goh, K. J., Ju, P. J. and Feng, Z. W. (2011). A Novel Role for Neural Cell Adhesion Molecule in Modulating Insulin Signaling and Adipocyte Differentiation of Mouse Mesenchymal Stem Cells. *Journal of Cell Science.* **124**, 2552-2560.

Yu, C., Takeda, M. and Soliven, B. (2000). Regulation of Cell Cycle Proteins by TNF-alpha and TGF-beta in Cells of Oligodendroglial Lineage. *Journal of Neuroimmunology.* **108**(11), 2-10.

Zhu, J., Lum, P. Y., Lamb, J., GuhaThakurta, D., Edwards, S. W., Thieringer, R., Berger, J. P., Wu, M. S., Thompson, J., Sachs, A. B. and Schadt, E. E. (2004). An Integrative Genomics Approach to the Reconstruction of Gene Networks in Segregating Populations. *Cytogenetic Genome Research.* **105**, 363-374.